Big Data and Modernizing Federal Statistics

Ron Jarmin Ph.D.
Assistant Director,
Research and Methodology Directorate

September 19, 2014

The Need to Modernize

- Reduce costs for legacy products
- Declining response rates and rising cost of direct data collection
- Competitive challenge – alternative information sources
  - But these sources may also be beneficial for federal statistics
- Demand for more timely and flexible data products
- Demand for information on small domains

How to Modernize

- Technology offers many opportunities – sometimes referred to as Big Data
- Data Rich World
- New ways to capture source data
- New ways to provide statistical information to users

How to Modernize
(continued)

- Many new sources and means of capturing raw source data for economic and social statistics:
  - New sources of administrative data (e.g., real estate records)
  - Social media
  - Sensor Data
  - Passive collections

How to Modernize
(continued)

- These newer sources free us from the constraints of paper survey forms
  - Flexibility to rapidly create new products
  - Unstructured (or less structured) data pose challenges

- Leverage statistical agencies' strengths
  - Lots of experience using large administrative datasets to augment survey collections
  - Future vision: use surveys to augment and add meaning and scientific value to non-survey data

Laying the Groundwork

- Need progress on methodological issues
  - What new data products are needed?
  - Modernize or discontinue existing products?
  - Need improved methods of estimation from unstructured, mashed-up data.
  - How to measure the quality of such statistics
  - Disclosure avoidance processes

Laying the Groundwork
(continued)

- Need progress on computational issues
  - Even larger volumes of data
  - Requires updated hardware and software
  - Can we support real or near real-time production of statistical information?
  - Cloud?
  - Where do source data live? Census (or contractor) controlled storage or at the source (federated)?
  - Upskilling staff in new techniques and data culture

Laying the Groundwork
(continued)

- Need progress on Policy issues
  - Privacy concerns
  - Data access issues – legal authorities and agreements
  - Risks versus rewards
  - Role of public vs private sector provision of economic and social statistics
  - Stakeholder buy-in

Laying the Groundwork
(continued)

- Progress needed on user engagement
  - New and modernized products meet users' needs
  - Transparency of sources and methods
  - Users understand and are able to draw correct inferences from and apply the new products

What are we doing now?

- External collaborations are key to success.
  - NSF Census Research Network
  - Other academic collaborations – e.g., MIT, Georgetown, VT, Stanford, Chicago, AIR
  - Private sector – e.g., Google, ESRI, UPS, First Data
  - Other agencies – e.g., FRB, NIST, USPTO
  - Other International Statistical Organizations and the United Nations
  - RDCs

What are we doing now?
(continued)

- Training existing staff
  - Census / U of Chicago Big Data Class
- Recruiting new staff with the right skill sets
- Projects
  - 2020 Administrative records and field reengineering
  - BDS/Patent data linkages
  - Pilot project with AIR and Big 10 universities – production function of federally funded research

Expected Outcomes

- Enhanced understanding of the cost/benefits and risk/rewards of more fully utilizing the Data-Rich environment to produce federal statistics
- Identification of opportunities to enhance source data and improve methods and processes to better meet users needs
- Enhanced external collaboration and staff development
- Modernization of hardware, software and personware
- Maintenance or enhancement of the role of federal statistical agencies and their data products